TSVWG
Internet-Draft
Intended status: Informational
Expires: December 27, 2019

F. Chen
W. Sun
X. Yu
Huawei Technologies Co., Ltd.
R. Even, Ed.
Huawei

June 25, 2019

Data Center Congestion Management requirements
draft-yueven-tsvwg-dccm-requirements-00

Abstract

On IP-routed datacenter networks, RDMA is deployed using RoCEv2 protocol. RoCEv2 specification does not define a strong congestion management mechanisms and load balancing methods. RoCEv2 relies on the existing Link-Layer Flow-Control IEEE 802.1Qbb(Priority-based Flow Control, PFC) to provide a lossless fabric. RoCEv2 Congestion Management(RCM) use ECN(Explicit Congestion Notification, defined in RFC3168) to signal the congestion to the destination and use the congestion notification to reduce the rate of injection and increase the injection rate when the extent of congestion decreases. More and more practice of congestion management for RoCEv2 appear in the industry, such as DCQCN(Data Center Quantized Congestion Notification). This document describes the current state of flow control and congestion handling in the DC using RoCEv2 and provides requirements for new directions for better congestion control.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on December 27, 2019.

Table of Contents

1. Introduction

With the emerging Distributed Storage, AI/HPC(High Performance Computing), Machine Learning, etc., modern datacenter applications demand high throughput(40Gbps and above) with ultra-low latency of

less than 10 microsecond per hop from the network, with low CPU overhead. The high link speed (>40Gb/s) in Data Centers (DC) are making network transfers complete faster and in fewer RTTs. Network traffic in a data center is often a mix of short and long flows, where the short flows require low latencies and the long flows require high throughputs.

On IP-routed datacenter networks, RDMA is deployed using RoCEv2 protocol. RoCEv2 [RoCEv2] is a straightforward extension of the RoCE protocol that involves a simple modification of the RoCE packet format. RoCEv2 packets carry an IP header which allows traversal of IP L3 Routers and a UDP header that serves as a stateless encapsulation layer for the RDMA Transport Protocol Packets over IP.

RoCEv2 Congestion Management (RCM) provides the capability to avoid congestion hot spots and optimize the throughput of the fabric. RCM relies on the existing Link-Layer Flow-Control IEEE 802.1Qbb(PFC) [IEEE.802.1QBB_2011] to provide a drop free network. RoCEv2 Congestion Management(RCM) also use ECN [RFC3168] to signal the congestion to the destination and use the congestion notification as an input to the sender to reduce the rate of injection and increase the injection rate when the extent of congestion decreases. The rate reduction by the sender as well as the increase in data injection is left to the implementation.

An enhancement to the congestion handling for ROCEv2 is the DCQCN [DCQCN] providing similar functionality to QCN and DCTCP, it is implemented in some of the ROCEv2 NICs but is not part of the ROCEv2 specification. As such, vendors have their own implementations which makes it difficult to interoperate with each other efficiently.

iWARP [RFC5040] provides a TCP based transport of RDMA, it is implemented in the NIC and is leveraging TCP retransmission and does not require a lossless fabric

2. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

3. Abbreviations

RCM - RoCEv2 Congestion Management

PFC - Priority-based Flow Control

ECN - Explicit Congestion Notification

DCQCN - Data Center Quantized Congestion Notification

AI/HPC - Artificial Intelligence/High-Performance computing

ECMP - Equal-Cost Multipath

NIC - Network Interface Card

4. Current Congestion Management mechanisms

4.1. Priority-based Flow Control (PFC)

RDMA can be deployed using the RoCEv2 protocol [RoCEv2], and relies on IEEE 802.1Qbb Priority-based Flow Control (PFC) [IEEE.802.1QBB_2011] to enable a drop-free network.

PFC is a link level protocol that allows a receiver to assert flow control by requesting the transmitter to pause sending traffic for a specified priority. However, because PFC will stop all traffic in a particular traffic class at the ingress port, the flows destined to other ports will also be blocked.

The known problems of PFC are head-of-line blocking, unfairness, deadlock [deadlocks]

4.2. Explicit Congestion Notification

Explicit Congestion Notification (ECN) [RFC3168] is used by the network to notify that congestion is detected before actually removing packets. Data Center TCP (DCTCP) [RFC8257]: TCP Congestion Control for Data Centers is an Informational RFC that extends the Explicit Congestion Notification (ECN) processing to estimate the fraction of bytes that encounter congestion, DCTCP then scales the TCP congestion window based on this estimate. DCTCP does not change the ECN reporting in TCP. Other ECN notification mechanisms for UDP based transports are specified for RTP in [RFC6679] and for QUIC [I-D.ietf-quic-transport]. The ECN notification are reported from the end receiver to the sender and the notification includes only the occurrence of ECN in the TCP case and the number of ECN marked packet for RTP and QUIC.

5. Congestion Management Practice

5.1. Packet Retransmission

NICs were not designed to deal with losses efficiently. Receiver discards out-of-order packets. Sender does go-back-N on detecting packet loss. RoCEv2 adopt Go-back-N loss recovery and needs lossless layer 2 (by using PFC) for good performance.

iWARP [RFC5040] provides a TCP based transport of RDMA, it is implemented in the NIC and is leveraging TCP retransmission and does not require a lossless fabric.

Based on iWARP congestion and packet loss handling an experiment to optimize the congestion control is in the improved RoCE NIC design [IRN] that makes two key changes to current RoCE NICs: (1) improving the loss recovery mechanism (similar to TCP with SACK), and (2) basic end-to-end flow control (termed BDP-FC) which bounds the number of in-flight packets by the bandwidth-delay product of the network, BDP-FC is a static value that is calculated based on the number of hops between the sender and the receiver. The tests results show that it provides better congestion handling comparing to DCQCN [DCQCN]. IRN work without PFC which is one of the concerns when using DCQCN.

Enhancements such as selective retransmission can be considered to not rely on a lossless network.

5.2. Congestion Control Mechanisms

5.2.1. RTT-based Congestion Control

The typical practice of RTT based Congestion Control is TIMELY [TIMELY]. TIMELY introduces the simple packet delay, measured as round-trip times at hosts, is an effective congestion signal without the need for switch feedback. TIMELY measures RTT with microsecond accuracy, and these RTTs are sufficient to estimate switch queuing. TIMELY can adjust transmission rates using RTT gradients to keep packet latency low while delivering high bandwidth. TIMELY is a delay-based congestion control protocol for use in the datacenter.

Because the RDMA transport is in the NIC and sensitive to packet drops, PFC is necessary because packets drops hurt performance badly. TIMELY needs PFC to provide lossless underlay network.

5.2.2. Credit-based Congestion Control

ExpressPass [ExpressPass] is an end-to-end credit-scheduled, delay-bounded congestion control for data centers. ExpressPass uses credit packets to control congestion even before sending data packets, which enables to achieve bounded delay and fast convergence. It uses end-

to-end credit transfer for bandwidth allocation and fine-grained packet scheduling.

5.2.3. ECN-based Congestion Control

Data Center Quantized Congestion Notification (DCQCN) [DCQCN] is an end-to-end congestion control scheme for RoCEv2. DCQCN is a combination of ECN and PFC to support end-to-end lossless Ethernet. The idea behind DCQCN is to allow ECN to do flow control by decreasing the transmission rate at the sender when congestion starts, thereby minimizing the time PFC is triggered. Configuring the ECN and PFC timeouts is challenging when there are more routers in the DC.

5.3. Re-ordering

When the packets arrive at the destination out-of-order, the destination should store the packets to restore the order. Destination should assign special buffer resource to perform re-ordering. There are many methods to implement the re-ordering either on the switches or on the NIC side.

5.4. Load Balancing

5.4.1. Equal-cost multi-path routing (ECMP)

RoCEv2 packets use an opaque flow identifier in the UDP Source Port field for ECMP method to implement path selection mechanisms for load balancing and improve utilization of the fabric topology. Traditional ECMP cannot balance loads well in the data center network because it splits loads at the granularity of flow. The finer the granularity of load balancing, the more effective the load balancing is and the higher the utilization of network bandwidth can be achieved.

5.4.2. Flowlet

The typical Flowlet-based load balancing is CONGA [CONGA]. CONGA is a network-based distributed congestion-aware load balancing mechanism for datacenters. It splits TCP flows into flowlets, estimates real-time congestion on fabric paths, and allocates flowlets to paths based on feedback from remote switches.

Flowlets are bursts of packets from a flow. The idle interval between two bursts of packets is larger than the maximum difference in latency among the paths. So the second burst can be sent along a different path than the first without reordering packets.

5.4.3. Per-packet

The effect of packet-based load balancing is the best because the corresponding granularity is the smallest. The consequence is that packets belonging to the same flow will be allocated to different paths. When the forwarding delays of paths are different, it is possible that packets may arrive at the receiver out-of-order.

6. Data Center Congestion Management requirements

The first issue is with incast traffic. Network congestion happens in the network routers when the incoming traffic is larger than the bandwidth of the outgoing link on which it has to be transmitted. Congestion is the primary source of loss in the network, congestion leads to performance degradation.

Another issue to address is packet loss due to out-of-order packets which may happen when load balancing is used. RoCEv2 adopt Go-back-N loss recovery and requires lossless fabric to prevent retransmission but is not addressing the packet loss due to re-ordering.

RoCEv2 relies on Link-Layer Flow-Control IEEE 802.1Qbb(PFC) [IEEE.802.1QBB_2011] to provide a lossless underlay networks. Lossless networks is implement by a mechanism of flow control, which pauses the traffic with priority granularity in the incoming link before the buffer overfills, and by that prevents the case of dropping packets [CongestionManagment]. However, PFC can lead to poor application performance due to problems like head-of-line blocking and unfairness [DCQCN].

Although DCQCN is widely deployed, due to the lack of formal specification, vendors have their own implementations which make it difficult to interoperate with each other efficiently. Moreover, the potential new congestion control mechanisms should also be considered to be compatible with existing ones.

Besides, with the development of RDMA fabric, the mixture of RDMA traffic and normal TCP traffic might also bring issues due to their employed different flow control and congestion control mechanisms.

In order to achieve the high throughput and low latency in the large-scale datacenter network, the following requirements for datacenter network congestion management are suggested:

- Resolve incast traffic in the network.
- Provide more efficient network congestion management for RDMA traffic to avoid retransmission.

- Provide better interoperability between vendors.
- Provide fairness mixture of RDMA traffic and normal TCP traffics.
- Provide compatibility when more than one congestion control mechanism is used.

7. Summary

As discussed in Section 6, we need an enhancement to current RDMA transport protocols with stronger capability of congestion management to achieve the high throughput and low latency in the large-scale datacenter network. The solution should also have more flexible requirement from the underlay network. The solution should work with ROCEv2 but should be more general so it can be used with iWARP as well.

8. Security Considerations

TBD

9. IANA Considerations

No IANA action

10. References

10.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

10.2. Informative References

[CONGA] Alizadeh, M., Edsall, T., Dharmapurikar, S., Vaidyanathan, R., Chu, K., Lam, V. T., Matus, F., Pan, R., Yadav, N., and G. Varghese, "CONGA: Distributed Congestion-Aware Load Balancing for Datacenters", 2 2015, <https://people.csail.mit.edu/alizadeh/papers/conga-sigcomm14.pdf>.

[CongestionManagment]
"Understanding RoCEv2 Congestion Management", 12 2018, <https://community.mellanox.com/s/article/understanding-rocev2-congestion-management>.

[DCQCN] Zhu, Y., Eran, H., Firestone, D., Guo, C., Lipshteyn, M., Liron, Y., Padhye, J., Raindel, S., Yahia, M. H., and M. Zhang, "Congestion control for large-scale RDMA deployments. In ACM SIGCOMM Computer Communication Review, Vol. 45. ACM, 523-536.", 8 2015, <https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf >.

[deadlocks]
Hu, S., Zhu, Y., Cheng, P., Guo, C., Tan, K., Padhye, J., and K. Chen, "Deadlocks in Datacenter Networks: Why Do They Form, and How to Avoid Them", 11 2016, <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/10/hotnets16-final67.pdf>.

[ExpressPass]
Cho, I., Han, D., and K. Jang, "ExpressPass: End-to-End Credit-based Congestion Control for Datacenters", 10 2016, <https://arxiv.org/pdf/1610.04688.pdf>.

[I-D.ietf-quic-transport]
Iyengar, J. and M. Thomson, "QUIC: A UDP-Based Multiplexed and Secure Transport", draft-ietf-quic-transport-20 (work in progress), April 2019.

[IEEE.802.1QBB_2011]
IEEE, "IEEE Standard for Local and metropolitan area networks--Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks--Amendment 17: Priority-based Flow Control", IEEE 802.1Qbb-2011, DOI 10.1109/ieeestd.2011.6032693, September 2011, <http://ieeexplore.ieee.org/servlet/opac?punumber=6032691>.

[IRN] Mittal, R., Shpiner, A., Panda, A., Zahavi, E., Krishnamurthy, A., Ratnasamy, S., and S. Shenker, "Revisiting Network Support for RDMA. In Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '18)", 8 2018, <https://dl.acm.org/citation.cfm?id=3230557>.

[RFC3168] Ramakrishnan, K., Floyd, S., and D. Black, "The Addition of Explicit Congestion Notification (ECN) to IP", RFC 3168, DOI 10.17487/RFC3168, September 2001, <https://www.rfc-editor.org/info/rfc3168>.

[RFC5040] Recio, R., Metzler, B., Culley, P., Hilland, J., and D. Garcia, "A Remote Direct Memory Access Protocol Specification", RFC 5040, DOI 10.17487/RFC5040, October 2007, <https://www.rfc-editor.org/info/rfc5040>.

[RFC6679] Westerlund, M., Johansson, I., Perkins, C., O'Hanlon, P., and K. Carlberg, "Explicit Congestion Notification (ECN) for RTP over UDP", RFC 6679, DOI 10.17487/RFC6679, August 2012, <https://www.rfc-editor.org/info/rfc6679>.

[RFC8257] Bensley, S., Thaler, D., Balasubramanian, P., Eggert, L., and G. Judd, "Data Center TCP (DCTCP): TCP Congestion Control for Data Centers", RFC 8257, DOI 10.17487/RFC8257, October 2017, <https://www.rfc-editor.org/info/rfc8257>.

[RoCEv2] "Infiniband Trade Association. Supplement to InfiniBand architecture specification volume 1 release 1.2.2 annex A17: RoCEv2 (IP routable RoCE).", <https://cw.infinibandta.org/document/dl/7781>.

[TIMELY] Mittal, R., Lam, T., Dukkipati, N., Blem, E., Wassel, H., Ghobadi, M., Vahdat, A., Wang, Y., Wetherall, D., and D. Zats, "RTT-based Congestion Control for the Datacenter", 8 2015, <https://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p537.pdf>.

Authors' Addresses

Fei Chen
Huawei Technologies Co., Ltd.

Email: chenfei57@huawei.com

Wenhao Sun
Huawei Technologies Co., Ltd.

Email: sam.sunwenhao@huawei.com

Xiang Yu
Huawei Technologies Co., Ltd.

Email: yolanda.yu@huawei.com

Roni Even (editor)
Huawei

Email: roni.even@huawei.com